Présentation du modèle de données de Crowdpac France

Le modèle de données de Crowdpac France analyse différentes sources d'informations publiques concernant les candidats potentiels et déclarés à l'élection présidentielle d'avril 2017, afin de donner des renseignements sur leurs positions et opinions politiques :

- *Discours et programmes politiques* – Le modèle de données de Crowdpac analyse ce que disent les candidats, ainsi que les mots et les phrases qu'ils utilisent le plus dans leurs programmes, leurs discours, leurs sites Web et leurs écrits, ainsi que sur les réseaux sociaux et dans les discours parlementaires. Combinées avec les positions clés définies dans leurs programmes, ces informations donnent une bonne indication de leurs priorités politiques.
- *Réseaux sociaux* – Notre algorithme analyse les relations entre plusieurs millions de comptes Twitter. Pour cela, il se base sur les réseaux d'abonnés de centaines de personnalités politiques. Cela nous donne une bonne indication de la position de ces personnalités en matière de politique et de leur réseau de soutien.
- *Votes* – Crowdpac a analysé les informations rendues publiques grâce aux initiatives d'accès libre aux données, et concernant les votes de l'Assemblée nationale sur la période de 2012 à 2016. Cette analyse nous a permis de générer un score relatif au vote pour les membres de l'Assemblée nationale.

Ces informations sont ensuite combinées afin d'affecter des scores idéologiques globaux à chaque candidat sur deux axes idéologiques déterminés via l'analyse des tendances clés identifiées dans les données :

1. Pour les questions de société, le spectre s'étend de « conservateur » à « progressiste ».
2. Pour les questions économiques, il va d'« interventionniste » à « libéral ».

Des informations supplémentaires concernant certains éléments clés du modèle de données de Crowdpac France sont fournies ci-dessous.

I. Textes et discours politiques

Crowdpac combine deux types d'analyse de texte afin d'évaluer les positions idéologiques des candidats à l'élection présidentielle française, sur la base de leurs textes et de leurs discours. Bien que mises en œuvre de façon différente, ces deux approches sont basées sur la notion selon laquelle les candidats

ayant des positions similaires s'expriment d'une façon présentant des similitudes quantifiables. Ces modèles sont combinés afin de générer des scores textuels pour chaque candidat. La page de profil de chaque candidat contient des nuages de mots clés présentant les mots et les phrases que ce candidat utilise le plus.

A. Collecte des données

En décembre 2016 et janvier 2017, nous avons utilisé des sources d'informations publiques pour constituer un corpus pour chaque candidat déclaré ou potentiel à l'élection présidentielle d'avril. Ces sources comprennent : les publications Facebook des candidats, les sites Web de leur campagne, leur blog personnel ou celui de leur parti (si le candidat en est l'unique représentant), les discours publics retranscrits à l'écrit, les discours parlementaires (Parlement européen et Sénat), les ouvrages publiés par ces candidats[1] et les hashtags mentionnés dans leurs tweets.[2] Dans tous les cas, nous avons cherché à utiliser uniquement les informations se rapportant à leur programme politique, et non des informations de type biographique ou d'une autre nature. Nous avons ensuite appliqué des procédures standard de nettoyage (recherche du radical, élimination des mots vides) afin de générer une matrice de fréquence pour les termes utilisés par chaque candidat. Chaque terme peut être un seul mot ou une expression contenant jusqu'à trois mots.

Début 2017, la taille des ensembles de données correspondant à chaque candidat était la suivante :

Nombre de tokens dans le corpus de chaque candidat

Montebourg	44940
Bayrou	23021
Hamon	32267
Dupont-Aignan	76017
De Rugy	55550
Fillon	55299
Hollande	142658
Bennahmias	92047
Juppé	110629
Le Pen	98565
Macron	94452
Mélenchon	492181
Arthaud	188619
Poutou	61410
Pinel	3899
Peillon	XXXX

[1] Nous avons inclus des textes issus de livres spécifiquement consacrés aux opinions ou programmes politiques de MM. Bayrou, De Rugy, Dupont-Aignan, Filoche, Fillon, Hamon, Jadot, Juppé, Macron, Mélenchon, Montebourg, Valls.
[2] Les tweets et les publications Facebook des candidats ont été fournis par make.org. Les tweets pris en compte datent de 2009 à la fin du mois d'octobre 2016.

Valls	130213
Jadot	178420

B. Méthode de création du dictionnaire

L'approche utilisée pour le dictionnaire suit les grandes lignes de la méthodologie décrite dans « Laver and Garry 2000 ».[3] Nous commençons par générer un dictionnaire de termes et d'expressions correspondant à une position sur deux dimensions : la politique économique (du pôle interventionniste au pôle libéral) et la politique sociale (du pôle conservateur au pôle progressiste). Ces termes ont été sélectionnés dans un vaste ensemble de mots et d'expressions qui, selon notre corpus, sont statistiquement les plus susceptibles d'être utilisés par les principaux partis politiques français. Nous avons sélectionné des mots et des phrases dont la simple mention traduit de façon non ambiguë une position particulière par rapport à l'une de ces dimensions, ou à ces deux dimensions, dans l'environnement politique actuel de la France. Exemples :

Démondialisation : Interventionniste
Entrepreneurial : Libéral
Émancipateur : Progressiste
Assimilation : Conservateur

Pour la plupart des candidats répertoriés ci-dessus, ce processus identifie des milliers d'utilisations de termes correspondant à une position idéologique définie sur l'un de nos deux axes. Ensuite, nous calculons le score *tf-idf* de tous les termes inclus dans le corpus de chaque candidat.[4] Cette méthode standard permet de mesurer la fréquence d'un terme dans un corpus donné par rapport à sa fréquence dans l'ensemble des textes. Nous calculons ensuite un score brut pour chaque candidat sur chaque dimension à l'aide de la formule suivante :

$$\text{score\_brut}_c = \frac{\text{abs}(\sum_{w \in \text{WR}} tf_w * i_w) * m_d}{\text{abs}(\sum_{w \in \text{WR}} tf_w * i_w) * m_d + \text{abs}(\sum_{w \in \text{WL}} tf_w * i_w)}$$

W_L et W_R correspondent respectivement aux ensembles de tokens relevant d'une idéologie de gauche (L = « left », gauche) et de droite (R = « right », droite) dans le corpus. tf_w est le score *tf-idf* du token *w* dans ce corpus. i_w est le libellé idéologique du token dans la plage [-2, 2]. m_d est le multiplicateur appliqué pour corriger l'écart qui se développe généralement dans le dictionnaire de chaque dimension concernant la distribution des mots associés à la gauche ou à la droite.

Ces scores sont ensuite normalisés par rapport à une plage prédéfinie correspondant à cette dimension, ce qui permet d'ajuster la distribution des mots dans l'échantillon de texte collecté. Nous avons également adapté ces résultats par rapport à la proportion de mots centristes (tels que « bipolarisation ») identifiés dans chaque dimension, afin de déceler les tentatives effectuées par certains

[3] Laver, Michael, et John Garry. « Estimating policy positions from political texts. » *American Journal of Political Science* (2000) : 619-634.
[4] Salton, Gerard et Michael J. McGill. « Introduction to modern information retrieval. » (1986).

candidats pour combler le clivage gauche-droite.

C. Modèle « Wordscores »

Nous combinons l'approche semi-supervisée ci-dessus, selon laquelle les connaissances antérieures sont appliquées pour attribuer des scores politiques à des mots et à des expressions, avec une approche supervisée qui permet d'évaluer les textes des candidats par rapport à une dimension donnée en se basant sur plusieurs textes de référence. Pour ce faire, nous avons utilisé l'algorithme *Wordscores*[5] précédemment utilisé avec succès pour l'analyse spatiale des manifestes des partis européens.

Wordscores se base sur un ensemble de textes de référence dont les positions par rapport au spectre politique analysé sont connues *a priori*. Par exemple, on pourra choisir le manifeste du Parti communiste comme texte de référence pour le côté gauche du spectre économique, et Les Républicains pour le côté droit. Pour calculer la position d'un nouveau texte par rapport à la dimension définie par les documents de référence, l'algorithme examine chaque mot qui s'y trouve et compare sa fréquence avec celle que l'on retrouve dans les textes de référence. Cela permet de calculer la probabilité conditionnelle selon laquelle le nouveau document pourrait être l'un des textes de référence. D'une certaine façon, l'ensemble de textes de référence (ou la matrice de fréquence des termes générée à partir de celui-ci) remplace le dictionnaire de codage utilisé dans l'approche décrite précédemment. En combinant ces probabilités, on obtient un score pour chaque corpus de la dimension prise en compte.

L'algorithme est appliqué à la même matrice de fréquence des termes que celle utilisée dans l'approche décrite à la section (B). Pour chacune des deux dimensions, nous avons sélectionné trois textes de référence de candidats définissant la gauche, la droite et le centre de l'échelle. Comme précédemment, nous appliquons une fonction de lissage aux scores bruts générés par *Wordscores* afin de compenser les écarts liés à la taille des textes collectés pour chaque candidat.

II. Analyse des réseaux d'abonnés Twitter

Les modèles précédents reposent sur les différences systématiques observées dans le vocabulaire utilisé par les candidats pour les situer par rapport à un espace politique. Notre second modèle, lui, analyse les réseaux d'abonnés Twitter et se base sur la tendance des individus à suivre les comptes de personnalités ou de partis politiques dont les positions coïncident globalement avec les leurs. En analysant ces réseaux à l'aide d'un modèle de correspondance, nous pouvons là encore générer un modèle spatial de préférences politiques, à la fois pour les comptes des abonnés et pour les comptes suivis.[6]

---

[5] Laver, Michael, Kenneth Benoit et John Garry. « Extracting policy positions from political texts using words as data. » *American Political Science Review* 97.02 (2003) : 311-331.

[6] Barberá, Pablo. « Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. » *Political Analysis* 23.1 (2015) : 76-91. Cet ouvrage est basé sur longues recherches dans le domaine de l'analyse spatiale des votes législatifs (Poole, Keith T. et Howard Rosenthal. « A spatial model for legislative roll call analysis. » *American Journal of Political Science* (1985) : 357-384.) et plus récemment, des dons faits

Nous avons collecté des données issues des réseaux d'abonnés Twitter d'environ 500 personnalités politiques françaises de premier plan. Le modèle crée un tableau de contingence n x m dans lequel chaque cellule contient 1 ou 0 pour indiquer si la personne A (dans l'ensemble de N abonnés collecté) suit la personne B (dans l'ensemble de M comptes politiques collectés). Après filtrage des comptes dont les réseaux sont insuffisants ou présentent trop d'irrégularités, nous appliquons un algorithme espérance-maximisation à la matrice de contingence restante afin d'estimer de façon itérative les points latents idéaux de tous les comptes.[7] Les scores bruts sont ensuite normalisés par rapport à la plage [-10 – 10] prédéterminée.

III. Relevés de votes

L'Assemblée Nationale française a rendu publics les relevés de votes de tous ses membres pour des milliers de votes sur la période de 2012 à 2016. Crowdpac a analysé ces relevés en se concentrant plus particulièrement sur les votes les plus controversés et significatifs marquant une rupture de la tendance à suivre la ligne du groupe ou du parti. L'objectif est de produire des scores de vote supplémentaires concernant les candidats à la présidentielle également membres de l'Assemblée nationale.

aux partis politiques (Bonica, Adam. « Mapping the ideological marketplace. » *American Journal of Political Science* 58.2 (2014) : 367-386.)

[7] Nous utilisons le modèle « Network Ideal Response Theory » inclus dans le package emIRT R, que nous complétons par des estimations des positions des différents partis, et par les positions de quelques candidats. Le modèle est appliqué au moyen de l'algorithme Wordfish décrit dans Slapin, Jonathan B. et Sven-Oliver Proksch. « A scaling model for estimating time-series party positions from texts. » *American Journal of Political Science* 52.3 (2008) : 705-722.